Chapter 1: Linear Regression Analysis

Dr. Abbas Rammal

Bachelor's degree in Mathematics
Option: DATA SCIENCE
September 2023

Plan

1. Introduction
2. Deterministic and Stochastic Relation
3. Deterministic relation
4. Stochastic relation
5. Least squares method
6. Explained and unexplained variation

Introduction

- **Purpose**: Regression analysis is the study of the stochastic dependence relationship that links two or more variables.

- **Application areas**:
  - Physics, chemistry, astronomy
  - Biology, medicine
  - Geography
  - Economy
  - Sociology...

- **Examples: Study the relationship between**
  - The distance traveled and the price of gasoline
  - Supply and demand
  - Sales and demand

Relation between two variables

- **Data**: Consider X and Y two variables.
- **Goal**: Study the relationship between Y and X.
- **Example**: Let X be the height (in cm) and Y the weight (in kg).

  How does a person's weight Y vary depending on their height X?

Practically:

1. We draw a sample of size n.
2. We measure the variables X and Y on this sample.
3. We obtain paired data

Indiv	Poids en kg(s)	Taille en cm(s)
1	74	180
2	80	184
3	62	165
4	78	184
5	75	170
6	58	163
7	50	155
8	85	178
9	70	175
10	60	168
MOY	68,20	172,20
E-T	10,64	9,14

Deterministic and Stochastic Relation

- The relationship between two variables X and Y can be:
  - exact ⇒ deterministic
  - not exact ⇒ stochastic

Deterministic Relation

- The relationship between two variables X and Y is an exact relationship.
- **Examples:**
  - X: amount in euros and Y: amount in dollar
  - X: distance and Y: ticket price.
  - X: temperature in Celsius and Y: temperature in Fahrenheit.

Deterministic relationship

The deterministic relation is of the form

$$Y = f(X)$$

where f is a determined function.

Deterministic and Stochastic Relation

- The function f can take different forms: linear, quadratic, exponential...
- We are interested in the linear relationship.
- The function f is then written in the following form:

$$f(x) = \beta_0 + \beta_1 x$$

where the parameters β_0 and β_1 are two fixed real numbers.

Examples:

1. Example euros in dollars: Y = 0.7579X2

2. Example of temperature from Celsius to Fahrenheit: $y = \frac{9}{5}x + 32$

- **Practically**:

1. We draw a sample of n data
2. We check that the data are aligned.
3. If this case is verified, then: the deterministic linear model.
4. If this case is not verified, then: the non-deterministic linear model

We must then look for the line that best fits the sample.

Stochastic Relation

- The relationship between X and Y is not exact ⇒ The linear relationship is not deterministic ⇒ The linear relationship is stochastic.
- **Example**: Let X be the height and Y the weight.
  - Several weights can correspond to 180 cm: 75 kg, 85 kg...
  - The data is no longer aligned.
  - For two identical weights, we have two different sizes.
- Model: In this case, the linear relationship between Y and X is written as:

$$f(x) = \beta_0 + \beta_1 x + \varepsilon$$

- The new variable ε represents individual behavior

- For each observation $i = 1 \ldots n$, we have:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

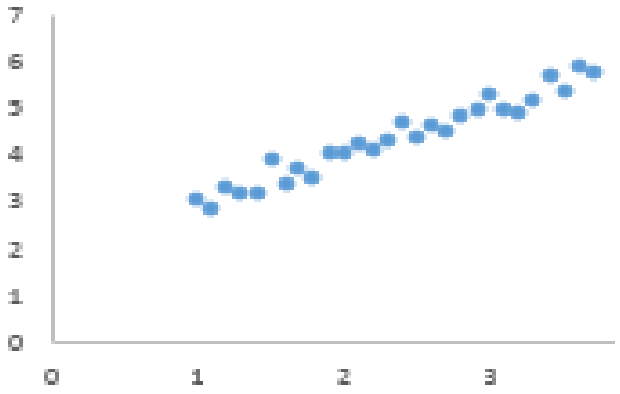
- The ε_i cannot be observed or calculated if we do not know the value of the parameters β_0 and β_1.
- The model is then a stochastic linear model or a linear regression model.

Data Representation: Scatter Plot

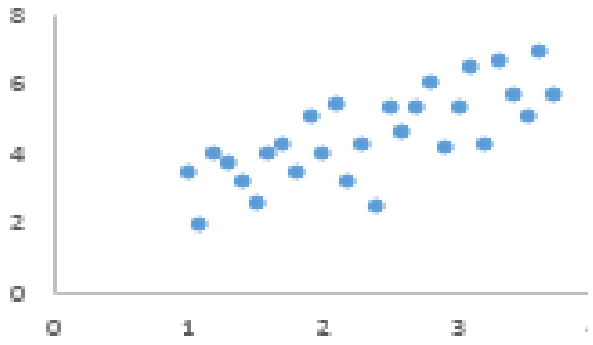
- The simplest method for observing the relationship between Y and X is to represent the pairs (x_i, y_i), for all observations, in a two-dimensional graph.
- This graph is called the scatter plot.
- From the scatter plot of points, we can observe the type of relationship existing between X and Y.

Data Representation: Scatter Plot

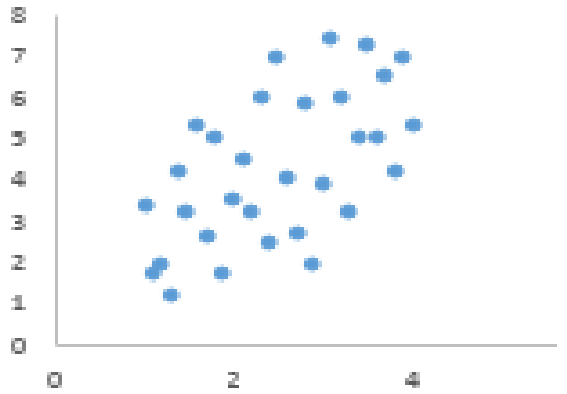
Strong linear correlation



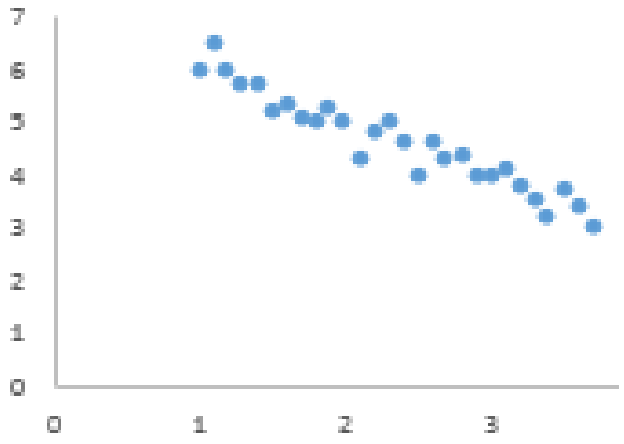
Moderate linear correlation



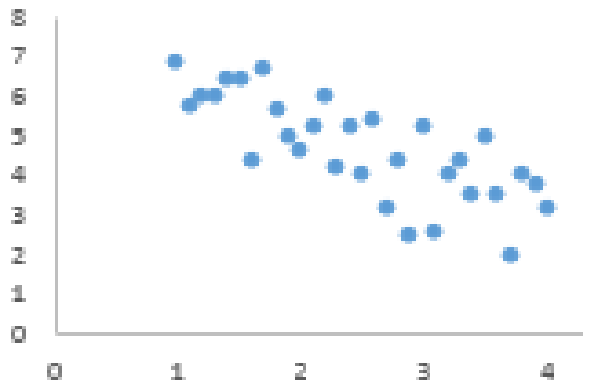
Weak linear correlation



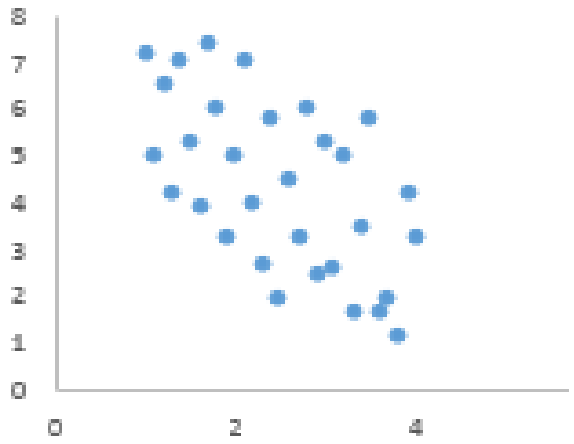
Strong linear correlation



Moderate linear correlation



Weak linear correlation



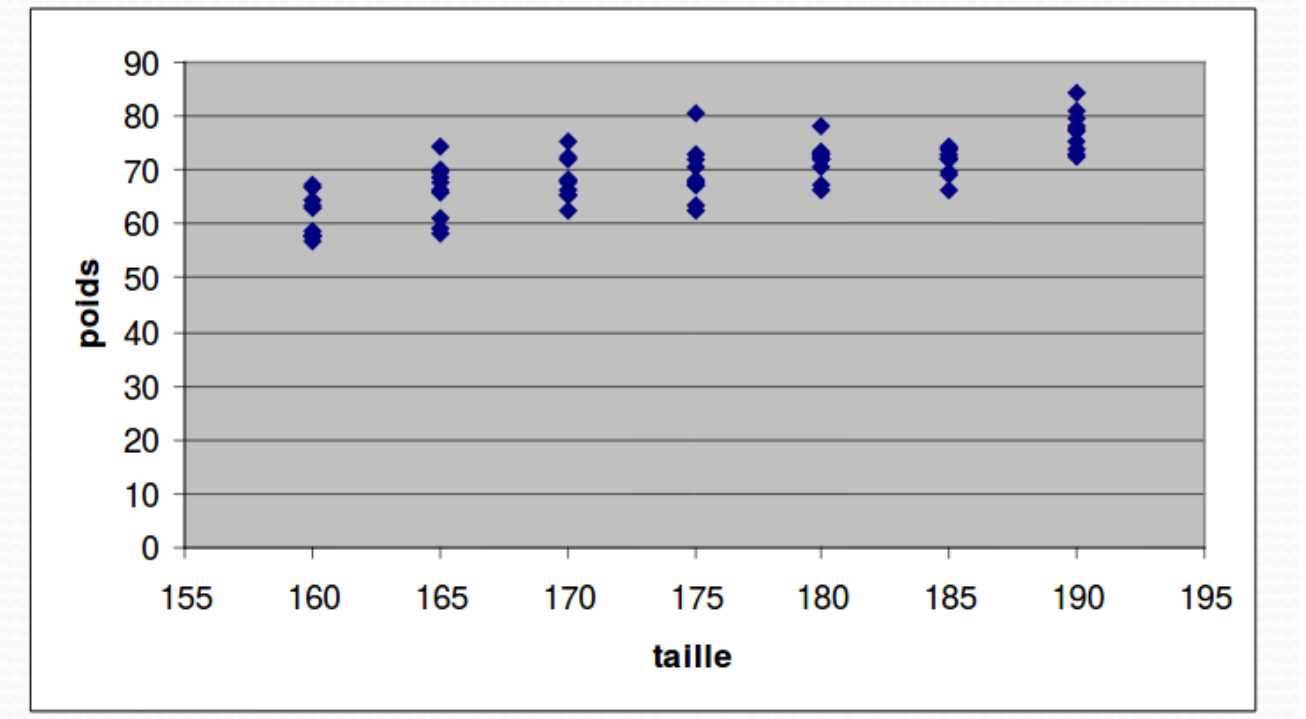
- Example:
- Variables: X: Height and Y: Weight.
- Sample: 70 individuals.

Obs. i	Taille x_i	Poids y_i	Obs. i	Taille x_i	Poids y_i	Obs. i	Taille x_i	Poids y_i
1	160	57,9	25	170	65,3	49	180	71,8
2	160	58,9	26	170	65,2	50	180	72,0
3	160	63,3	27	170	68,3	51	185	74,1
4	160	56,8	28	170	62,3	52	185	74,4
5	160	66,8	29	170	67,8	53	185	72,0
6	160	64,5	30	170	66,5	54	185	66,1
7	160	67,1	31	175	67,4	55	185	69,4
8	160	58,0	32	175	67,7	56	185	71,8
9	160	62,9	33	175	62,6	57	185	73,8
10	160	57,7	34	175	70,6	58	185	69,1
11	165	68,5	35	175	72,0	59	185	72,3
12	165	69,8	36	175	68,3	60	185	72,8
13	165	58,5	37	175	72,9	61	190	72,6
14	165	66,3	38	175	63,4	62	190	81,1
15	165	65,8	39	175	80,7	63	190	78,3
16	165	61,0	40	175	67,3	64	190	72,9
17	165	74,5	41	180	67,4	65	190	79,6
18	165	59,3	42	180	70,6	66	190	77,1
19	165	67,8	43	180	72,4	67	190	84,5
20	165	70,1	44	180	73,2	68	190	74,0
21	170	72,7	45	180	72,8	69	190	77,5
22	170	75,1	46	180	66,4	70	190	75,2
23	170	68,0	47	180	73,0			
24	170	72,2	48	180	78,0			

- Data representation:

- Comments:

1. Several Y for the same value of X. ⇒ Inadequate deterministic linear model.
2. When X increases, Y increases ⇒ Possible stochastic linear model.



- We note $\mu_y(x) =$ the average of y measured on all individuals for which x.
- The linear regression model is sometimes formulated in the form:

$$\mu_y(x) = \beta_0 + \beta_1 x$$

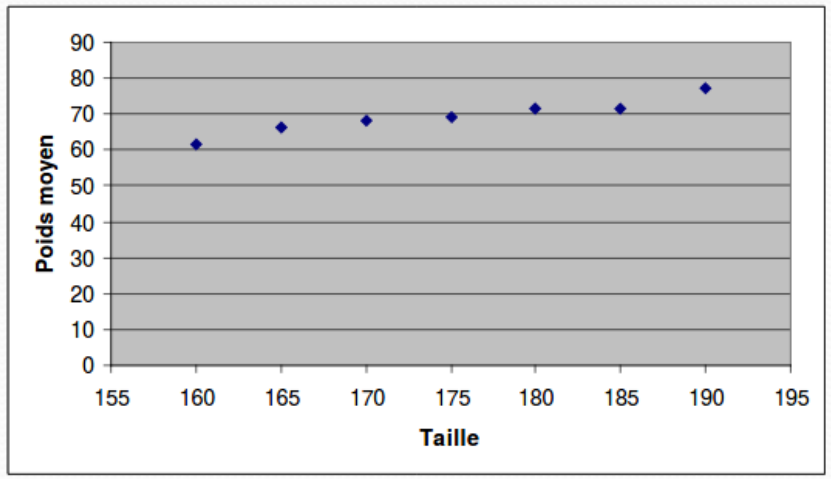
- Like ε, $\mu_y(x)$ is neither observable nor calculable.
- To calculate $\mu_y(x)$, it would be necessary to list all the individuals in the population.

- Practically: We estimate the theoretical average $\mu_y(x)$ by the empirical average of Y defined by

$$\bar{y}(x) = \frac{1}{n}\sum_{i=1}^{n} y_i(x)$$

- Example: For x = 160, we have: $\bar{y}(160) = \frac{57.6 + 58.9 + \cdots + 62.9 + 57.7}{10} = 61.39$

Taille	Poids
160	61,39
165	66,16
170	68,34
175	69,29
180	71,76
185	71,58
190	77,28



- These averages are approximately aligned on a straight line.
- This line is called: the regression line.
- It expresses the average of Y as a function of the different values of X.
- X explains Y ⇒ X is an independent (or explanatory) variable and Y is a dependent (or explained) variable.

- **Problem of regression analysis :** the regression line is unknown ⇒ the problem of regression analysis is to estimate β_0 and β_1 from a sample of data.
- **Choice of parameters**: Determine the line that best fits the data ⇒ $\hat{\beta}_0$ and $\hat{\beta}_1$ are estimators of β_0 and β_1.
- For $x = x_i$, the value of y calculated by the equation $\hat{y}(x) = \hat{\beta}_0 + \hat{\beta}_1 x$ is given by $\hat{y}_i(x) = \hat{\beta}_0 + \hat{\beta}_1 x_i$
- $\hat{y}_i$ is called the value estimated by the model.
- The $\hat{y}_i$ values make it possible to estimate unobservable quantities:

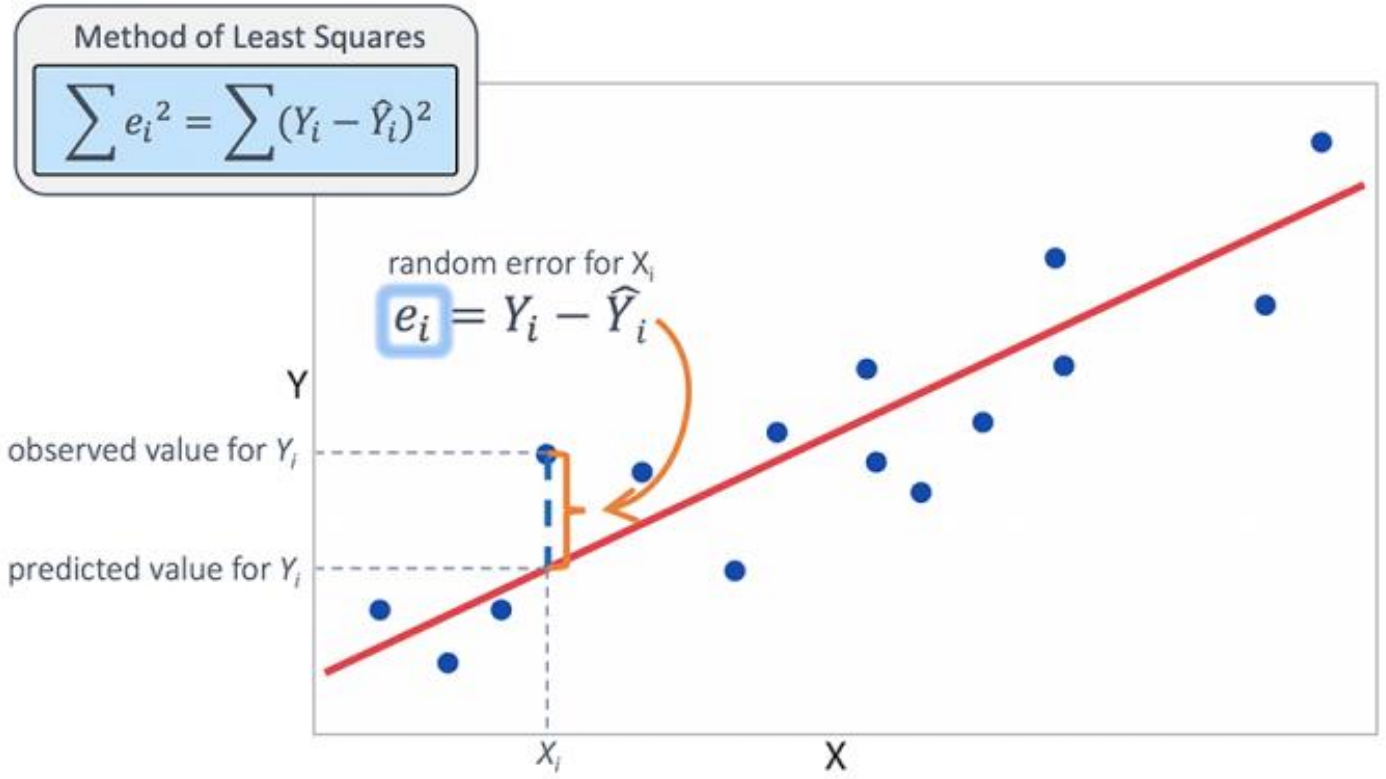
$$\varepsilon_i = y_i - \beta_0 - \beta_1 x_i$$

by the observable quantities:

$$e_i = y_i - \underbrace{(\hat{\beta}_0 + \hat{\beta}_1 x_i)}_{\hat{y}_i} = y_i - \hat{y}_i$$

- The quantities e_i are called the residuals of the model.

Illustration:



least squares method

- Most estimation methods consist of estimating the regression line by a line that minimizes a residual function.
- The best known is the least squares method.
- To estimate β_0 and β_1, we can use the least squares method which requires minimizing the sum of the squares of the residuals:

$$E = f(\beta_0, \beta_1) = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2.$$

- We must therefore solve the following optimization problem:

$$(\hat{\beta}_0; \hat{\beta}_1) = \underset{\beta_0, \beta_1}{argmin} E$$

- The partial derivative of E with respect to β_0 and β_1 are given by:

$$\frac{\partial E}{\partial \beta_0} = -2\sum(y_i - \beta_0 - \beta_1 x_i)$$

$$\frac{\partial E}{\partial \beta_1} = -2\sum x_i(y_i - \beta_0 - \beta_1 x_i)$$

- The estimators $\hat{\beta}_0$ and $\hat{\beta}_1$ are the solutions of the system of equations:

$$\sum(y_i - \beta_0 - \beta_1 x_i) = 0 \Rightarrow \sum y_i = n\hat{\beta}_0 + \hat{\beta}_1 \sum x_i \qquad (1)$$

$$\sum x_i(y_i - \beta_0 - \beta_1 x_i) = 0 \Rightarrow \sum x_i y_i = \hat{\beta}_0 \sum x_i + \hat{\beta}_1 \sum x_i^2 \qquad (2)$$

- Like $\bar{x} = \frac{\sum x_i}{n}$ and $\bar{y} = \frac{\sum y_i}{n}$ we obtain from equation (1):

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

- By replacing $\hat{\beta}_0$ in equation (2), we obtain:

$$\begin{aligned} \hat{\beta}_1 \sum x_i^2 &= \sum x_i y_i - \hat{\beta}_0 n\bar{x} \\ &= \sum x_i y_i - (\bar{y} - \hat{\beta}_1 \bar{x}) n\bar{x} \\ &= \sum x_i y_i - n\bar{x}\bar{y} + \hat{\beta}_1 n\bar{x}^2 \end{aligned}$$

- **Solution of the optimization problem**: We then obtain:

$$\begin{cases} \hat{\beta}_1 = \dfrac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2} = \dfrac{\sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \dfrac{\sum_{i=1}^{n} x_i(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \\ \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \end{cases}$$

Equation of the regression line

- The equation of the regression line is given by $\hat{y}(x) = \hat{\beta}_0 + \hat{\beta}_1 x$
- By definition of $\hat{\beta}_0$, the regression line always passes through the point $(\bar{x}, \bar{y})$.
- Note: The sum of the residuals is zero.

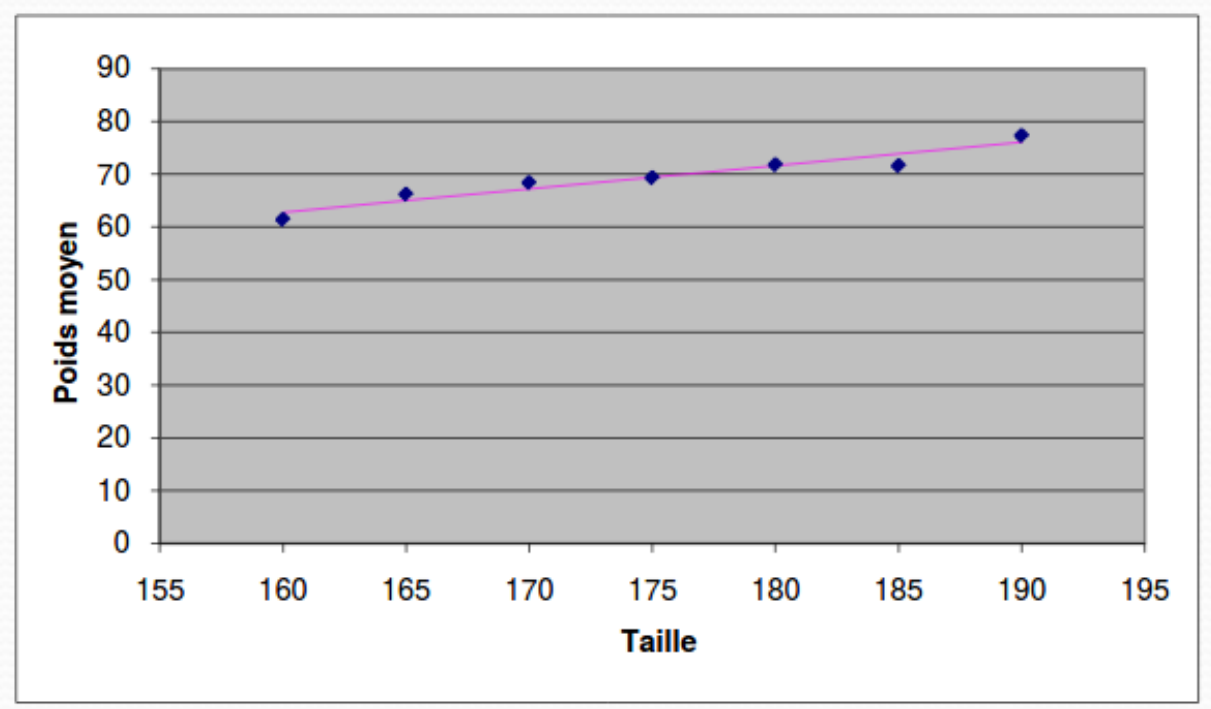
$$\sum e_i = \sum (y_i - \bar{y}) - \hat{\beta}_1 \sum (x_i - \bar{x})$$

$$\sum e_i = \underbrace{\sum (y_i - \bar{y})}_{0} - \hat{\beta}_1 \underbrace{\sum (x_i - \bar{x})}_{0} = 0$$

$$\sum y_i = \sum \hat{y}_i$$

- The mean of the y_i is equal to the mean of the $\hat{y}_i$

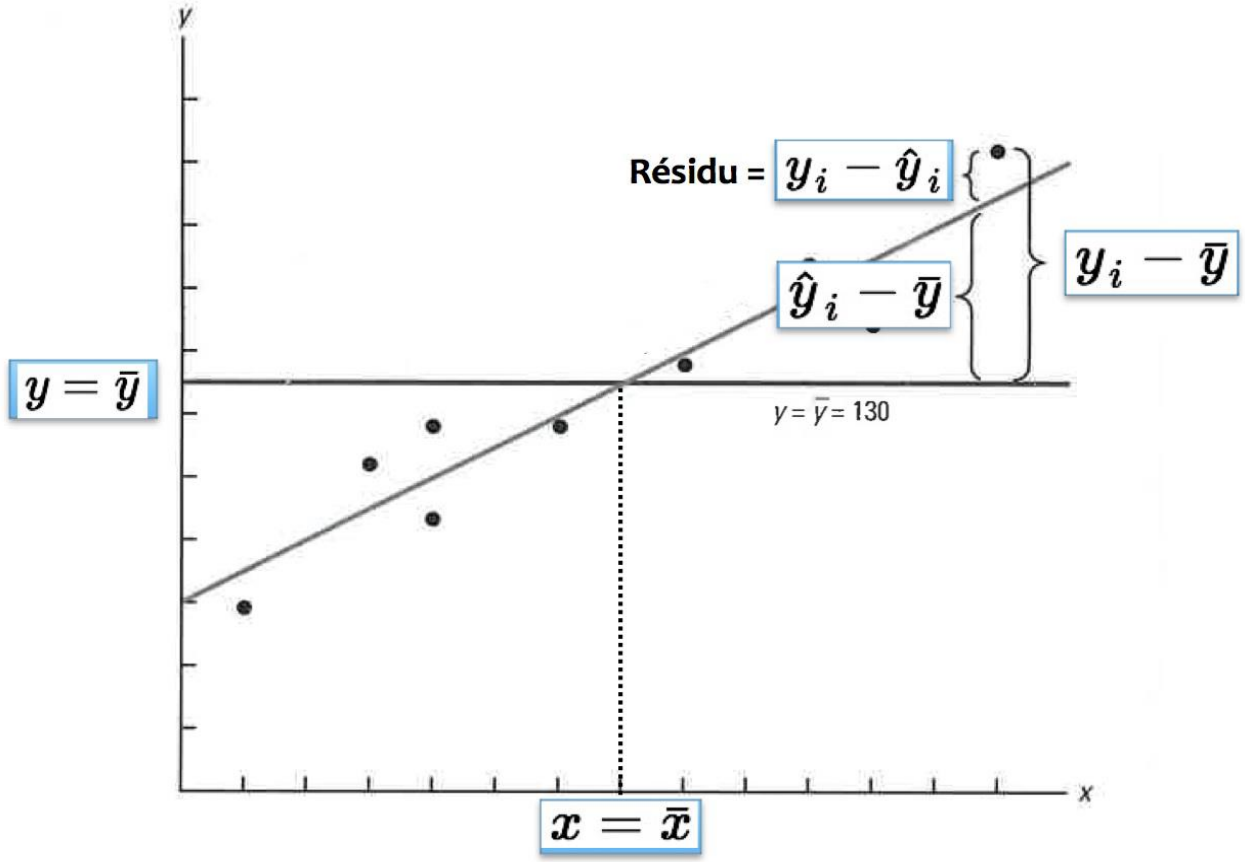
- Example: Coefficients of the regression line for the weight and height example:
- The slope: $\hat{\beta}_1 = 0.442$ and the intercept $\hat{\beta}_0 = -8.012$.



Explained and unexplained variation

- The goal of a linear regression model is to explain part of the variation of the explained variable Y due to its dependence on the explanatory variable X.
- If Y depends on X ⇒ if X varies then Y varies accordingly ⇒ Variation explained by the model.
- If we measure Y on individuals with the same value of X we can observe a certain variation on Y ⇒ Variation unexplained by the model.

- Variation situation: Total variation of Y = Explained variation + Unexplained variation.
- Measurement of the variation of Y: To measure the variation of Y, we must use the differences between the observations y_i and the mean $\bar{y}$.



- Decomposition of variation:

$$(y_i - \bar{y}) = \underbrace{(\hat{y}_i - \bar{y})} + \underbrace{(y_i - \hat{y}_i)}$$

- Why the least squares method? It maintains such a decomposition by considering the sum of the squares of these differences:

$$\sum (y_i - \bar{y})^2 = \sum (\hat{y}_i - \bar{y})^2 + \sum (y_i - \hat{y}_i)^2$$

Propriétés:

1. $\sum \hat{y}_i^2 = \sum \hat{y}_i y_i$
2. $\sum (y_i - \hat{y}_i)^2 = \sum y_i^2 - \sum \hat{y}_i^2$
3. $\sum (\hat{y}_i - \bar{y})^2 = \sum \hat{y}_i^2 - n\bar{y}^2 = \hat{\beta}_1^2 \sum (x_i - \bar{x})^2.$

- Notations:
- Total sum of squares: $SC_{tot} = \sum(y_i - \bar{y})^2$
- Sum of squares due to regression: $SC_{reg} = \sum(\hat{y}_i - \bar{y})^2$
- Sum of squares of residuals: $SC_{res} = \sum(y_i - \hat{y}_i)^2$

$$SC_{tot} = SC_{reg} + SC_{res}$$

- **Measurement of the percentage of the total variation explained by the model:**

  Introduction of a coefficient of determination, denoted R^2

$$R^2 = \frac{\text{Variation expliquée}}{\text{Variation totale}} = \frac{SC_{reg}}{SC_{tot}}$$

$$\begin{aligned} R^2 &= \frac{\sum(\hat{y}_i - \bar{y})^2}{\sum(y_i - \bar{y})^2} \\ &= \frac{\sum \hat{y}_i^2 - n\bar{y}^2}{\sum y_i^2 - n\bar{y}^2} \\ &= \hat{\beta}_1^2 \frac{\sum(x_i - \bar{x})^2}{\sum(y_i - \bar{y})^2} \\ &= \hat{\beta}_1^2 \frac{\sum x_i^2 - n\bar{x}^2}{\sum y_i^2 - n\bar{y}^2} \end{aligned}$$

- R^2 is between 0 and 1.
- R^2 = 1: case where the data are perfectly aligned (as is the case for a deterministic model).
- R^2 = 0: case where the variation of Y is not due to the variation of X. The data are not aligned at all.
- The closer R^2 is to 1, the more the data are aligned with the regression line.